AD-A192 648

2003Ol29OO4

~~ATION PAGE~~ — ~~READ INSTRUCTIONS BEFORE COMPLETING FORM~~

	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER

~~IME FILE COPY~~

4. TITLE (and Subtitle)

Nucleotide Sequence of the Protective Antigen
Gene of _Bacillus_ _anthracis_

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

S. L. Welkos, J. R. Lowe, F. Eden-McCutchan,
M. Vodkin, S. M. Leppla, J. J. Schmidt

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS

United States Army Medical Research Institute
  of Infectious Diseases
Fort Detrick, Frederick, MD 21701-5011

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Research and Development Command
Fort Detrick
Frederick, MD 21701-5011

12. REPORT DATE

2 February 1988

13. NUMBER OF PAGES

34 pages

14. MONITORING AGENCY NAME & ADDRESS(If different from Controlling Office)

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

DTIC
ELECTE
MAR 1 5 1988
S ⊂⊂H

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, If different from Report)

18. SUPPLEMENTARY NOTES

Submitted to Journal of Bacteriology, 1913 I Street, N.W., Washington, DC 20006

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

B. anthracis, protective antigen, gene, nucleotide sequence

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The DNA sequence of the protective antigen gene from _Bacillus_ _anthracis_ and the
5' and 3' flanking sequences were determined. Protective antigen is one of
three proteins comprising anthrax toxin. The open reading frame is 2319 base
pairs (bp) long, of which 2205 bp encode the 735 amino acids of the secreted
protein. This region is preceded by 29 codons, which appear to encode a signal
peptide having characteristics in common with those of other secreted proteins.
A consensus TATAAT sequence was located at the putative -10 promoter site. A
Shine-Dalgarno site similar to that found in genes of other Bacillus sp. (cont)

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE

20. ABSTRACT (cont)

was located seven bp upstream of the ATG initiation codon. The codon usage for the protective antigen gene reflected the high A + T (69%) base composition. The TAA translation stop codon was followed by an inverted repeat forming a potential termination signal. In addition, a 192-codon open reading frame of unknown significance, theoretically encoding a 21.6 kilodalton protein, preceded the 5' end of the protective antigen gene.

Nucleotide Sequence of the Protective Antigen Gene of

Bacillus anthracis

S. L. WELKOS[1*],J. R. LOWE[1] F. EDEN-MCCUTCHAN[2], M. VODKIN[1+],

S. H. LEPPLA[1], J. J. SCHMIDT[1]

*U.S. Army Medical Research Institute of Infectious Diseases[1],
Frederick, Maryland 21701-5011, and Molecular Biology and
Recombinant DNA Laboratories[2], Bionetics Research, Inc.,
Rockville, MD 20850*

Running title: Sequence of Protective Antigen Gene

Clearance date -

*Corresponding author
Telephone: 301/663-7341
+ Present address - Department of Veterinary Pathology,
University of Illinois, Urbana, IL 61801

88 3 14 057

ABSTRACT

The DNA sequence of the protective antigen gene from
Bacillus anthracis and the 5′ and 3′ flanking sequences were
determined. Protective antigen is one of three proteins
comprising anthrax toxin. The open reading frame is 2319 base
pairs (bp) long, of which 2205 bp encode the 735 amino acids of
the secreted protein. This region is preceded by 29 codons,
which appear to encode a signal peptide having characteristics in
common with those of other secreted proteins. A consensus TATAAT
sequence was located at the putative -10 promoter site. A
Shine-Dalgarno site similar to that found in genes of other
Bacillus sp. was located seven bp upstream of the ATG initiation
codon. The codon usage for the protective antigen gene reflected
the high A + T (69%) base composition. The TAA translation stop
codon was followed by an inverted repeat forming a potential
termination signal. In addition, a 192-codon open reading frame
of unknown significance, theoretically encoding a 21.6 kilodalton
protein, preceded the 5′ end of the protective antigen gene.

Bacillus anthracis is an important pathogen of animals and of people exposed to infected animals or their products. It can cause cutaneous anthrax, gastrointestinal anthrax, and an often fatal systemic pulmonary form of the disease (13, 21, 22). The two major virulence factors of *B. anthracis* are a poly-D-glutamic acid capsule and "anthrax toxin." DNA functions controlling toxin and capsule production are carried on *B. anthracis* plasmids pXO1 and pXO2, respectively (10, 31). The toxin is composed of three separate proteins, protective antigen (PA), edema factor (EF), and lethal factor (LF). The three proteins are nontoxic alone. However, PA in combination with LF causes death in rats (2), and PA combined with EF produces edema in the skin of guinea pigs and rabbits (21, 22). In addition to mediating the toxic effects of LF and EF, protective antigen induces immunity to infection and is the major component of the currently licensed human vaccine (13, 14, 21, 23, 41).

In order to understand the role of PA in the pathogenesis of disease and the induction of protective immunity, the DNA encoding PA has been cloned and sequenced. All three of the toxin proteins are encoded by the 176-kilobase pair (kb) plasmid pXO1 (31, 35, 42). Vodkin and Leppla (42) first reported the cloning of the PA gene in *Escherichia coli*. The gene was contained in a 6-kb *Bam*HI fragment of pXO1 cloned into plasmid pBR322. Full-size, biologically active PA was produced. The *Bacillus* promoter was present but expression of the gene by the recombinant plasmid (pSE36) in *E. coli* was low.

In a recent study, we subcloned the 6-kb insert of pSE36 into the plasmid vector pUB110 and transformed *B. subtilis* with the recombinant DNA (14). Two recombinants were isolated which produced large amounts of full-size PA despite the presence of deletions in the 6-Kb insert of approximately 2.7 kb and 3.4 kb, respectively. In vitro concentrations of PA produced by the recombinants were similar or greater than those observed with *B. anthracis* (14). Protective antigen, a protein of approximately 85 kilodaltons (kDa) by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (14, 21, 42), requires a coding region of 2 - 2.5 kb.

The purpose of the present study was to map and sequence the coding region of PA. Partial digestion and religation of plasmid pSE36 (which has the 6-kb insert) yielded a smaller derivative plasmid, pPA26, which contains a 4.2-kb insert encoding full-size PA. In this report, the nucleotide sequence of this insert and analysis of the PA coding region are presented.

MATERIALS AND METHODS

Bacteria and plasmids. Isolates of *E. coli* K12 strain HB101, transformed with pSE36 or pPA26, were the sources of plasmid DNA; and strain JM103 (29) was used to propagate M13 phage derivatives.

Subcloning and detection of PA-producing recombinants. The isolation of recombinant *E. coli* (pSE36) has been described (42). Briefly, pSE36 consists of plasmid pBR322 with a 6-kb *Bam*HI fragment encoding the PA protein from plasmid pXO1 of *B. anthracis*. To obtain derivatives having smaller insert DNA, plasmid pSE36 DNA was partially digested with *Hin*dIII and religated. *E. coli* strain HB101 was transformed with the plasmid DNA, and recombinants were tested for the presence of the PA gene by immunological assay (42). The site and biological activity of PA produced by the recombinants were tested by a Western blot procedure and the CHO cell elongation assay, respectively (14, 20, 42).

Isolation of DNA. Plasmid pPA26 DNA was prepared from cleared lysates by ultracentrifugation in cesium chloride/ethidium bromide gradients according to methods described by Maniatis et al (25). The DNA was digested simultaneously with *Hin*dIII and *Bam*HI, and the 4.2-kb insert encoding PA was isolated as a 2.2-kb *Hin*dIII and 2.0-kb *Hin*dIII-*Bam*HI fragment (Fig. 1). The DNA fragments were purified by preparative gel electrophoresis, the bands excised, and the DNA extracted with phenol for cloning in M13.

Nucleotide sequence analysis. The two fragments were each cloned into phages M13mp10 and M13mp11, and the dideoxy chain termination method (30, 38) was used to sequence the DNA. Initially, data were collected by using the universal primer (Pharmacia P-L Biochemicals, Piscataway, N.J.). Using these data, we synthesized oligonucleotide primers 18 nucleotides long to collect each additional data segment (37). The oligonucleotides were prepared by the phosphoramidite method (Applied Biosystems, Foster City, Calif.). The products of the sequencing reactions were separated in 7% denaturing polyacrylamide gels, and data read from the autoradiograms were compiled and melded by using the GEL program in the IntelliGenetics Molecular Biology software package (IntelliGenetics, Inc., Mountain View, Calif.).

Enzymes and reagents. Restriction endonucleases were purchased from International Biotechnolgies, Inc. (New Haven, Conn.) and Bethesda Research Laboratories (Gaithersburg, Md.) and were used as recommended by the suppliers. T4 DNA ligase and deoxynucleoside and dideoxynucleoside triphosphates were from Pharmacia P-L. Klenow fragment was purchased from Boehringer Mannheim Biochemicals (Indianapolis, Ind.), and α-^{32}P-deoxynucleoside triphosphates (300 - 800 curies/mmole, 11.1 - 29.6 TBq/mmole) were from Amersham (Arlington Heights, Ill.).

Computer analysis of DNA sequence and protein secondary structure. The sequence in pPA26 of *B. anthracis* DNA was analyzed by several computer software packages. The MOLGENJR programs (J. R. Lowe, Fed. Proc. 45:1582, 1986) were run on an

IBM-PC microcomputer to confirm experimental restriction enzyme cleavage patterns, deduce open reading frames, and translate the primary DNA sequence into amino acid sequences. Other programs from the MOLGENJR package were used to examine the translated peptide sequences, calculate codon usage and polypeptide molecular weights, and plot hydropathy and secondary structure histograms. We used a VAX 750 minicomputer, executing program SEQ in the IntelliGenetics Molecular Biology software package, to search for regions of dyad symmetry and to calculate free energies of base pairing in potential DNA hairpin secondary structure. Other unpublished programs and algorithms were used to search for potential activation sequences (ENHANCE2.MSB) and significant open reading frames (ORFREAL.MSB) and to create condensed, dot-matrix hydropathy and secondary structure histograms (AGNAKDCF.MSB). These are available from J.R. Lowe.

RESULTS AND DISCUSSION

Cloning of the PA gene and sequencing strategy. The
Protective Antigen gene of *B. anthracis* was originally cloned
into pBR322 as a 6-kb insert (42). Digestion and religation of
the recombinant plasmid pSE36 yielded a smaller plasmid with a
4.2-kb insert (pPA26), which retained the PA gene (Figure 1A).
E. coli transformants of pSE36 and pPA26 both produced proteins
of about 83 kDa on SDS-PAGE which reacted specifically with
anti-PA antibody on Western blot analysis and were biologically
active in the CHO cell elongation assay (20; data not shown). To
determine the location and direction of transcription of the PA
gene, the 4.2-kb insert was excised, digested with *Hind*III and
*Bam*HI into two fragments of 2.0-kb and 2.2-kb, and sequenced as
indicated in Figure 1B and 1C.

Nucleotide sequence analysis - PA.

(i) **Open reading frame.** The nucleotide sequence of the PA
protein is shown in Figure 2. Analysis of the sequence revealed
an open reading frame 2319 base pairs (bp) long. The structural
gene for the mature protein began at nucleotide 1891, coding for
a glutamic acid residue, and the translated sequence was in
agreement with both the N-terminal amino acid sequence and the
amino acid composition determined previously. The coding region
for this portion of the PA gene was 2205 bp long, encoding a 735-
amino acid protein with a theoretical molecular weight of 82,684
daltons. The size of the mature PA protein as determined by
sequence analysis was similar to that estimated by SDS-PAGE

analysis of PA from *Bacillus* culture supernatants, 83 - 85 kDa (14, 21, 42). The final residue of the coding region (glycine) was followed by the consensus TAA stop codon (nucleotide 4096). Thus, as indicated in Figure 1C, all except the N-terminal 53 amino acids were encoded within the 2.0-kb *Hind*III-*Bam*HI fragment at the 3' end of the 4.2-kb *B. anthracis* insert. This location of the gene at the end of the insert confirms the position of the PA gene mapped in recently isolated *B. subtilis* recombinants (14). In that study, cloning of the *B. anthracis* insert into *B. subtilis* (pUB110) yielded two plasmid recombinants with deletions at the 5' end of the insert. The smaller recombinant plasmid retained just 2.6-kb of DNA at the 3' end of the PA insert but produced full-sized, functional PA (14).

Preceding the sequence encoding the 83 kDa PA protein (starting at nucleotide 1891) were two ATG codons in phase with the open reading frame, at nucleotides 1834 and 1804. Similar to other *Bacillus* proteins, PA is a secreted protein and is probably synthesized as a precursor having a signal peptide. The methionine codon at nucleotide 1804 appears to be the likely starting point for translation. It would initiate a sequence having several characteristics in common with other *Bacillus* signal sequences that have been identified. The 29-residue peptide that would be encoded is typical of the size of other *Bacillus* signal sequences (3, 17, 19, 31, 44, 46). Also, the positively charged, N-terminal five amino acids (Met-Lys-Lys-Arg-Lys), the hydrophobic central region (residues 6

to 21), the terminal alanine residue are characteristic of bacterial signal peptides (3, 17, 19, 31, 44, 46).

(ii) **Transcription and translation regulatory regions.** A putative Shine-Dalgarno ribosomal binding site, indicated in Figure 2, is located seven bp upstream of the ATG codon at nucleotide 1804. The sequence of this site, AAACGAG, and the distance separating it from the initiation codon, closely resemble the characteristics of the Shine-Dalgarno sites reported for several other *Bacillus sp* genes (5, 26, 33, 43, 46). The Shine-Dalgarno sequence has a calculated binding energy with *B. subtilis* 16S rRNA of -14.0 kcal/mole (1, 27, 40). Possible promotor sequences are underlined in Figure 2. The putative RNA polymerase recognition site (TATAAT) at nucleotide 1764 is identical to the *E. coli* and *B. subtilis* σ^{43} -10 consensus sequences. The 6-base sequence starting at nucleotide 1738, and separated by 20 bp from the -10 site, resembles the conserved -35 site of *E. coli* and the -35 site reported for genes of gram positive organisms (36). The optimal distance between the -10 and -35 RNA polymerase recognition regions in *B. anthracis* genes is unknown. In *E. coli*, these sequences are separated by 16 to 19 bp, with 17 being the most frequent and resulting in maximal promoter strength (36). *Bacillus* promoters, especially those recognized by σ^{43}-containing RNA polymerase, are often similar in their sequence and spacing to *E. coli* promoters; however several different promoter sequences have been identified (8, 15, 43, 44). Also, distances between the two promoter regions as long as

21 bp have been reported for other sequences, *e.g.* the pertussis toxin gene (24). In vitro and in vivo transcription analyses will be necessary to locate the precise promoter region for the PA gene. An inverted repeat forming a potential termination structure was located 3' of the translation stop codon as shown in Figure 3A. The putative hairpin structure contained 19 complementary nucleotide pairs and two T-G mismatches between nucleotides 4142 and 4188. The structure had a strong predicted free energy of base-pair formation (ΔG_f = -22.2 kcal/mole).

We observed three additional regions forming potential stem-and-loop structures, which showed significant probabilities and negative free energies of formation. The sequence from nucleotides 868 to 926 (Fig. 3B), was inside the 192-codon open reading frame and had a strong calculated ΔG_f of -25.4 kcal/mole. The second region of dyad symmetry (Fig. 3C), from nucleotides 1263 to 1346, had a predicted ΔG_f = -19.6 kcal/mole. The third region (Fig. 3D) spanned the PA promoter from nucleotides 1722 to 1779 and had a predicted ΔG_f = -15.8 kcal/mole. If any or all of these regions is recognized as a transcriptional terminator in *E. coli*, their presence could possibly explain the low PA expression from the original clones (5 - 10 ng PA/ml) (42).

(iii) **Base composition and codon usage**. The base composition of the coding strand of the PA gene was: A = 39%, T = 30% (A + T = 69% of total), G = 17%, C = 14% (G + T = 31%). The codon usage is shown in Table 1. There was a preference for A

and T at the third position in the codons, which might reflect the high A + T content. The codon usage was similar to that of another *Bacillus* gene of plasmid origin, the crystal protein toxin gene of the related species *B. thuringiensis* (39). A major difference between the two codon profiles was the lack of cysteine residues in PA. The codon usage in the PA gene differed from that in genes for toxins and other proteins produced by other gram-positive and gram-negative bacteria (Table 1 and data not shown).

Analysis of protein structure from the nucleotide sequence. The prediction of the amino acid sequence of PA and the deduction of protein structural information were performed by algorithms of the computer programs described above (J. R. Lowe, Fed. Proc. **45**:1582, 1986 and other unpublished programs). The algorithms used to predict the hydropathic profile and the protein secondary structure are based on the methods of Kyte and Doolittle (18) and Chou and Fasman (4), respectively. These predictions are shown in Figure 4. The amino-terminal portion of the putative signal peptide is hydrophilic, whereas the central core was hydrophobic, as expected from comparisons with similar analyses of other proteins with confirmed signal sequences (data not shown).

Regions of the sequence upstream of the PA gene. Other open reading frames, in addition to the longest one of 2319 bp encoding PA, were found in the 4.2-kb sequence. The only open reading frame at least 100 codons long was a 576-nucleotide sequence (ORF1) beginning with an ATG at position 416 upstream of

the PA gene. The 192-codon open reading frame encodes a polypeptide with a calculated Mr of 21,610 daltons. The codon usage of the translated region is similar to that observed for the PA gene (Table 1). A computer analysis (ORFREAL.MSB) of this open reading frame according to the method of Fickett (7) calculated a 92% coding probability. A similar analysis of the PA coding region also gave a 92% coding probability. Potential -10 and -35 RNA polymerase recognition sites, but no consensus Shine-Dalgarno site, on the 5' side of the cryptic open reading frame were identified. The open reading frame terminated with a TAG stop codon. Figure 5 is a hydropathy plot and secondary structure analysis of the putative protein. The sequence does not appear to encode a signal peptide but does have an interesting carboxy terminus rich in hydrophobic residues embedded in a region with a high probability for β-sheet structure. This suggests that the protein could be membrane-bound at its carboxy terminus. The significance of this putative gene is unknown and awaits analysis of expression experiments using the cloned plasmid DNA.

The availability of the complete nucleotide sequence of the PA of *B. anthracis* will serve several useful purposes. For example, the promoter sequence of PA can be probed by promoter-proving vectors and the sequence altered by site-specific mutagenesis. Thus, enhanced production of cloned PA in the *B. subtilis* or *E. coli* hosts will become feasible. Also, specific mutagenesis of the PA coding region could be done

to: (1) produce immunogenic, biologically inactive cross-reactive proteins for vaccine studies (13, 23, 41); or (2) examine the role of different domains of PA on binding to target cell membranes and to the EF and LF components of anthrax toxin (9, 20, 32). Finally, segments of the PA nucleotide sequence will be used as probes to examine the genetic organization of PA in variant strains of *B. anthracis* (23,41).

LITERATURE CITED

1. **Band, L., and D. J. Henner.** 1984. *Bacillus subtilis* requires a "stringent" Shine-Dalgarno region for gene expression. DNA **3**:17-21.

2. **Beall, F. A., M. J. Taylor, and C. B. Thorne.** 1962. Rapid lethal effect in rats of a third component found upon fractionating the toxin of *Bacillus anthracis*. J. Bacteriol. **83**:1274-1280.

3. **Blobel, G., F. Walter, C. N. Chang, B. K. Goldman, A. H. Erikson, and V. R. Lingappa.** 1979. Translocation of proteins across membranes: the signal hypothesis and beyond. Symp. Soc. Exp. Biol. **33**:9-36.

4. **Chou, P. Y., and G. D. Fasman.** 1978. Prediction of the secondary structure of proteins from their amino acid sequence, p. 45-148. *In* A. Meister (ed.), Advances in enzymology and related areas of molecular biology, vol. 47, John Wiley & Sons, New York.

5. **Duvall, E. J., D. M. Williams, S. Mongkolsuk, and P. S. Lovett.** 1984. Regulatory regions that control expression of two chloramphenicol-inducible *cat* genes cloned in *Bacillus subtilis*. J. Bacteriol. **158**:784-790.

6. **Fairweather, N. F., V. A. Lyness, D. J. Pickard, G. Allen, and R. O. Thomson.** 1986. Cloning, nucleotide sequencing, and expression of tetanus toxin fragment C in *Escherichia coli*. J. Bacteriol. **165**:21-27.

7. **Pickett, J. W.** 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. **10**:5303-5318.

8. **Fliss, E. R., and P. Setlow.** 1984. Complete nucleotide sequence and start sites for transcription and translation of the *Bacillus megaterium* protein C gene. J. Bacteriol. **158**:809-813.

9. **Friedlander, A. M.** 1986. Macrophages are sensitive to anthrax lethal toxin through an acid-dependent process. J. Biol. Chem. **261**:7123-7126.

10. **Green, B. D., L. Battisti, T. M. Koehler, C. B. Thorne, and B. E. Ivins.** 1985. Demonstration of a capsule plasmid in *Bacillus anthracis*. Infect. Immun. **49**:291-297.

11. **Greenfield, L., M. J. Bjorn, G. Horn, D. Fong, G. A. Buck, R. J. Collier, and D. A. Kaplan.** 1983. Nucleotide sequence of the structural gene for diphtheria toxin carried by corynebacteriophage β. Proc. Natl. Acad. Sci. U.S.A. **80**:6853-6857.

12. **Grosjean, H., and W. Fiers.** 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18**:199-209.

13. **Hambleton, P., J. A. Carman, and J. Melling.** 1984. Anthrax: the disease in relation to vaccines. Vaccine. **2**:125-132.

14. **Ivins, B. E., and S. L. Welkos.** 1986. Cloning and expression of the *Bacillus anthracis* protective antigen gene in *Bacillus subtilis*. Infect. Immun. **54**:537-542.

15. **Johnson, W., K. Moran, and R. Losick**. 1983. Two RNA polymerase sigma factors from *Bacillus subtilis* discriminate between overlapping promoters for a developmentally regulated gene. Nature (London) **302**:800-804.

16. **Jones, C. L., and S. A. Khan**. 1986. Nucleotide sequence of the enterotoxin B gene from *Staphylococcus aureus*. J. Bacteriol. **166**:19-33.

17. **Krail, G.** 1981. Transfer of proteins across membranes. Ann. Rev. Biochem. **50**:317-348.

18. **Kyte, J., and R. F. Doolittle**. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105-132.

19. **Lampen, J. O., and J. B. K. Nielsen**. 1982. Penicillinase and the Secretion of Proteins by *Bacilli*, p. 99-109. *In* A. T. Ganesan, S. Chang, and J. A. Hoch (ed.), Molecular cloning and gene regulation in *Bacilli*, Academic Press, New York.

20. **Leppla, S. H.** 1984. *Bacillus anthracis* calmodulin-dependent adenylate cyclase: chemical and enzymatic properties and interactions with eucaryotic cells, p. 189-198. *In* P. Greengard and G. A. Robinson (ed.), Advances in cyclic nucleotide and protein phosphorylation research, Raven Press, New York.

21. **Leppla, S. H., B. E. Ivins, and J. W. Exzell**. 1985. Anthrax toxin, p. 63-66. *In* L. Leive (ed.), Microbiology - 1985. American Society for Microbiology, Washington, D.C.

22. **Lincoln, R. E., and D. C. Fish**. 1970. Anthrax toxin, p. 361-414. *In* T. C. Montie, S. Kadis, and S. I. Ajl (ed.), Microbial toxins, vol. 3. Academic Press, Inc., New York.

23. **Little, S. F., and G. B. Knudson**. 1986. Comparative efficacy of *Bacillus anthracis* live spore vaccine and protective antigen vaccine against anthrax in the guinea pig. Infect. Immun. **52**:509-512.

24. **Locht, C., and J. M. Keith**. 1986. Pertussis toxin gene: nucleotide sequence and genetic organization. Science **232**:1258-1264.

25. **Maniatis, T., E. F. Fritsch, and J. Sambrook**. 1982. Molecular cloning. A laboratory manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

26. **McLaughlin, J. R., S. -Y. Chang, and S. Chang**. 1982. Transcriptional analyses of the *Bacillus licheniformis penP* gene. Nucleic Acids Res. **10**:3905-3919.

27. **McLaughlin, J., C. Murray, and J. Rabinowitz**. 1981. Unique features in the ribosomal binding site sequence of the gram-positive *Staphylococcus au 'us* B-lactamase gene. J. Biol. Chem. **256**:11283-11291.

28. **Mekalanos, J. J., D. J. Swartz, G. D. N. Pearson, N. Harford, F. Groyne, and M. de Wilde**. 1983. Cholera toxin genes: nucleotide sequence, deletion analysis and vaccine development. Nature (London) **306**:551-557.

29. **Messing, J.** 1983. New M13 vectors for cloning. Meths. Enzymol. **101**:20-78.

30. **Massing, J., R. Crea, and P. Seeburg**. 1981. A system for shotgun DNA sequencing. Nucleic Acids Res. **9**:309-321.

31. **Mikesell, O. P., B. E. Ivins, J. D. Ristroph, and T. M. Dreier**. 1983. Evidence for plasmid-mediated toxin production in *Bacillus anthracis*. Infect. Immun. **39**:371-376.

32. **O'Brien, J. O., A. Friedlander, T. Drier, J. Ezzell, and S. Leppla**. 1985. Effects of anthrax toxin components on human neutrophils. Infect. Immun. **47**:306-310.

33. **Ohmura, K., H. Yamazaki, Y. Takeichi, A. Nakayama, K. Otozai, K. Yamane, M. Yamasaki, and G. Tamura**. 1983. Nucleotide sequence of the promoter and NH$_2$-terminal signal peptide region of *Bacillus subtilis* α-amylase gene cloned in pUB110. Biochem. Biophys. Res. Commun. **112**:678-683.

34. **Palva, I., R. F. Pettarsson, N. Kalkkinen, P. Lahtovaara, M. Sarvas, H. Soderlund, K. Takkinen, and L. Kaariainen**. 1981. Nucleotide sequence of the promoter and NH$_2$-terminal signal peptide region of the α-amylase gene from *Bacillus amyloliquifaciens*. Gene **15**:43-51.

35. **Robertson, D. L. and S. H. Leppla**. 1986. Molecular cloning and expression in *E. coli* of the lethal factor gene of *Bacillus anthracis*. Gene **44**:71-78.

36. **Rosenberg, M. and D. Count**. 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. **13**:319-353.

37. **Sanchez-Pescador, R., and M. Urdea**. 1984. Use of unpurified

synthetic oligodeoxyribonucleotide primers for rapid
deoxynucleotide chain termination sequencing. DNA 3:339-343.

38. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA
sequencing with chain-terminating inhibitors. Proc. Natl.
Acad. Sci. U.S.A. **74**:5463-5467.

39. **Schnepf, H. E., H. C. Wong, and H. R. Whiteley.** 1985. The
amino acid sequence of a crystal protein from *Bacillus
thuringinensis* deduced from the DNA base sequence. J. Biol.
Chem. **260**: 6264-6272.

40. **Tinoco, I., P. N. Borer, B. Dangler, M. D. Levine, O. C.
Uhlenbeck, D. M. Crothers, and J. Gralla.** 1973. Improved
estimation of secondary structure in ribonucleic acids.
Nature (London) **246**:40-41.

41. **Turnbull, P.C. B., M. G. Broster, J.A. Carman, R. J.
Manchee, and J. Melling.** 1986. Development of antibodies to
protective antigen and lethal factor components of anthrax
toxin in humans and guinea pigs and their relevance to
protective immunity. Infect. Immun. **52**:356-363.

42. **Vodkin, M. H. and S. H. Leppla.** 1983. Cloning of the
protective antigen gene of *Bacillus anthracis*. Cell **34**:693-
697.

43. **Waalwijk, C., A. M. Dullemans, M. E. S. van Workum, and B.
Visser.** 1985. Molecular cloning and the nucleotide sequence
of the M$_r$ 28,000 crystal protein gene of *Bacillus
thuringiensis* subsp. *israelensis*. Nucleic Acids Res.
13:8207-8217.

44. **Wells, J. A., E. Ferrari, D. J. Henner, D. A. Estell, and E.Y. Chen**. 1983. Cloning, sequencing, and secretion of *Bacillus amyloliquefaciens* subtilisin in *Bacillus subtilis*. Nucleic Acids Res. **11**:7911-7925.

45. **Yamamoto, T., T. Tamura, and T. Yokota**. 1984. Primary structure of heat-labile enterotoxin produced by *Escherichia coli* pathogenic for humans. J. Biol. Chem. **259**:5037-5044.

46. **Yang, M., A. Galizzi, and D. Henner**. 1983. Nucleotide sequence of the amylase gene from *Bacillus subtilis*. Nucleic Acids Res. **11**:237-249.

FIGURE LEGENDS

FIG. 1 - Construction of plasmid containing the PA gene and sequencing strategy.

(A) Plasmid pPA26 was constructed from pSE36 by partial HindIII digestion. The 4.2-kb HindIII-BamHI portion of the plasmid (open box) contains the PA gene. The distances (in kb) between the BamHi (B) and HindIII (H) sites are indicated; EcoRI sites (E) are included.

(B) To sequence this insert, pPA26 was digested with BamHI and HindIII. The 2.2-kb HindIII and 2.0-kb HindIII-BamHI fragments were isolated and cloned into M13 mp10 and mp11. (C) The arrows indicate the direction and extent of sequencing of the DNA fragments, totalling 4235-kb. The hatched bar indicates the structural gene for the mature PA protein.

FIG. 2 - Nucleotide and amino acid sequence of the PA gene and 5' and 3' flanking sequences. The sequence shown corresponds to nucleotides 1 - 4235 on the map in Fig. 1. Restriction endonuclease sites described in Fig. 1 and in the text are indicated. The presumptive -35 and -10 sequences, and Shine-Dalgarno ribosomal binding site (rbs) of the PA gene and of the potential 192 nucleotide open reading frame are underlined, as are the translation start (ATG) and stop (TAG, TAA) codons. Arrows above the nucleotide sequence indicate initiation of translation of the potential open reading frame upstream of the PA gene (ORF1) and of the signal sequence (SIG) and mature

protein (MAT) of the PA gene. The 29-residue signal peptide is underlined. The translated amino acid sequences of the 192 nucleotide open reading frame and the PA gene, only, are shown. The potential stem-loop termination structure flanking the 3' end of the PA gene, and the three palindromic sequences on the 5' side of the PA gene are indicated by dashed lines between outward pointing arrowheads above the sequences.

FIG. 3 - Possible stem and loop structures found in the upstream and downstream sequences from the PA gene a_d in the putative peptide coding region. Numbering corresponds with that of Fig. 2. The calculated free energies of these conformations were (A) ΔG_f = -22.2 kcal/mole, (B) ΔG_f = -25.4 kcal/mole, (C) ΔG_f = -19.6 kcal/mole, and (D) ΔG_f = -15.8 kcal/mole.

FIG. 4 - Combination hydropathy/secondary structure plot of PA. The left margin contains the amino acid sequence of PA. Residue numbers are scaled on the right ordinate. The abscissa units are hydropathy values. Dot positions on the left portion of the plot indicate the most probable secondary structure feature predicted. Headings are H for helix, S for β-sheet, T for β-turn, and R for random coil. Region A is the hydrophobic signal sequence with its highly charged amino terminus. Regions B-J are potential antigenic sites in the sequence. Dot-matrix output was computer-generated with program AGNAKDCF.EXE. Algorithms developed

according to the schemes of Kyte and Doolittle ('8) and Chou and Fasman (4).

FIG. 5 - Combination hydropathy/secondary structure plot of the putative peptide. Plot organization and generation are the same as described in Figure 4. Region A is a long, highly hydrophobic 28-residue sequence with mostly β-sheet structure predicted. Regions B-D are potential antigenic sites in the sequence.

(A)

pSE36
(10.4 Kb)

EH B
.35
.03   1.8
H
4.0   2.2
H
2.0
B   E

Hind III
partial   Ligate

PPA26
(8.25 Kb)

EH
H
E
B

(B)
Hind III +
Bam HI

(C)

5' H   2187   2872   4235 3'
                                          H   E   B

200
bp
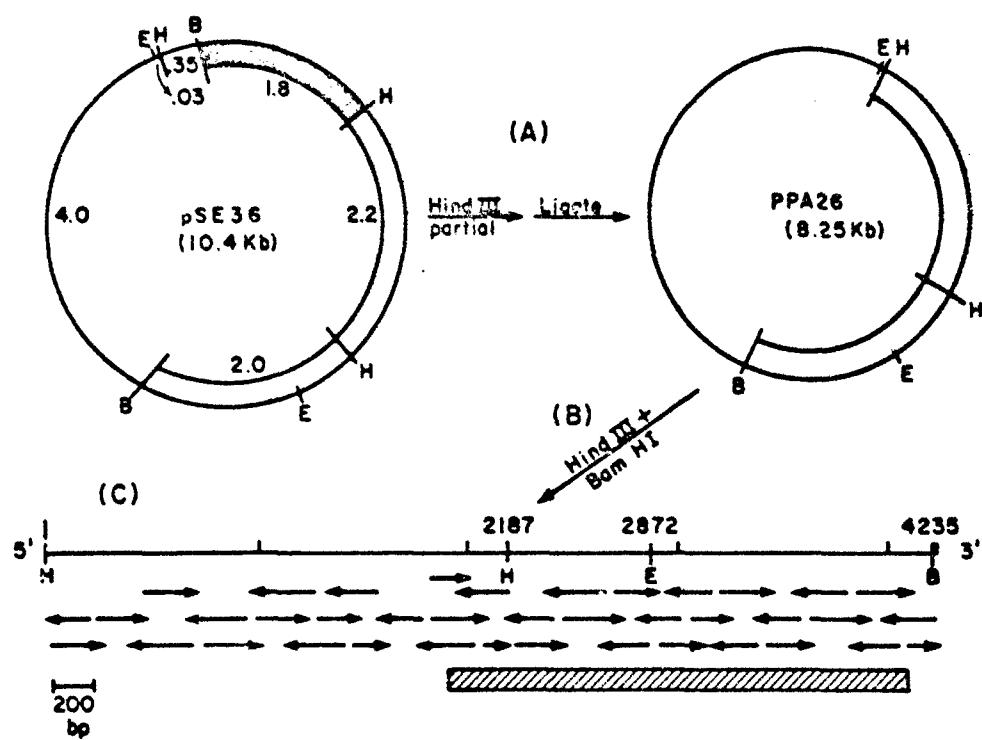
_Hind_III
        50   100
```
AAGCTTCTGTCATTCGTAAATTTCAAATAGAACGTAAATTTAGACTTCTCATCATTAAAATGAAAAATCTTATCTTTTTGATTCTATTGTATATTTTA
                                              150                               200
TTAAGGTGTTTAATAGTTAGAAAAGACAGTTGATGCTATTACTCCAGATAAAATATAGCTAACCATAAATTTATTAAAGAAACCTTGTTGTTCTAAATAA
                                              250                               300
TGATTTTGTGGATTCCGGAAATAGATACTGGTGAGTTAGCTCTAAATTTTATAGTGATTAACTAACAATTTATAAAGCAGCATAATTCAAATTTTTTAATT
                              350   -35    .          .        .  -10   .        400
GATTTTTCCTGAAGCATAGTATAAAAGAGTCAAGGTCTTCTAGACTTGACTCTTGGAATCATTAGGAATTAACAATATATATAATGCGCTAGACAGAATC
    ----->  ORF                               450                               500
AAATTAAATGCAAAAATGAATATTTTAGTAAGAGATCCATATCATTATGATAATAACGGTAATATTGTAGGGGTTGATGATTCATATTTAAAAAACGCAT
         MetAsnIleLeuValArgAspProTyrHisTyrAspAsnAsnGlyAsnIleValGlyValAspAspSerTyrLeuLysAsnAlaTyr
                                              550                               600
ATAAGCAAATACTTAATTGGTCAAGCGATGCAGTTTCTTTAAATCTAGATGAAGATGTAAATCAAGCACTATCTGGATATATGCTTCAAATAAAAAAACC
LysGlnIleLeuAsnTrpSerSerAspGlyValSerLeuAsnLeuAspGluAspValAsnGlnAlaLeuSerGlyTyrMetLeuGlnIleLysLysPro
                                              650                               700
TTCAAACCACCTAACAAACAGCCCAGTTACAATTACATTAGCAGGCAGGACAGTGGTGTTGGAGAATTGTATAGAGTATTATCAGATGGAGCAGGATTC
SerAsnHisLeuThrAsnSerProValThrIleThrLeuAlaGlyLysAspSerGlyValGlyGluLeuTyrArgValLeuSerAspGlyAlaGlyPhe
                                              750                               800
CTGGATTTCAATAAGTTTGATGAAAATTGGCGATCATTAGTAGATCCTGGTGATGATGTTTATGTGTATGCTGTTACTAAAGAAGATTTTAATGCAGTTA
LeuAspPheAsnLysPheAspGluAsnTrpArgSerLeuValAspProGlyAspAspValTyrValTyrAlaValThrLysGluAspPheAsnAlaValThr
                                    850        <-----------------------  --
CTCGAGATGAAAATGGTAATATAGCGAATAAATTAAAAAAACACCTTAGTTTTTATCGGTAAAATAAAAGAAATAAACATAAAACTACAAATATTAATAT
ArgAspGluAsnGlyAsnIleAlaAsnLysLeuLysAsnThrLeuValLeuSerGlyLysIleLysGluIleAsnIleLysThrThrAsnIleAsnIle
-------------------------->        .          .  950         .          .        1000
ATTTGTAGTTTTTATGTTTATTATATACCTCCTATTTATATTATTAGTAGCACAGTTTTTGCAAATCATGTAATTGTATACTTATCTATGTAGAGGTAT
PheValValPheMetPheIleIleTyrLeuLeuPheTyrIleIleSerSerThrValPheAlaAsnHisValIleValTyrLeuSerMet
                                              1050                              1100
CACAACTTATGAATAGTGTATTTTATTGAACGTTGGTTAGCTTGGACAGTTGTATGGATATGCATACTTTATAACGTATAAAAATTTCACGCACCACAATA
                                              1150                              1200
AAACTAATTTAACAAAAACAAAAACACACCTAAGATCATTCAGTTCTTTTAATAAGGAGCTGCCCACCTAGCTAAACCTAAATAATCTTTGTTTCACATA
                              1250        .     <----------------------------------
AGGTTTTTTTTCTAAATATACAGTGTAAGTTATTGTGAATTTAACCAGTATATATTAAAAATGTTTTATGTTAACAAATTAAATTGTAAAACCCCTCTTAA
-------------------------------------------->1350                              1400
GCATAGTTAAGAGGGGTAGGTTTTAAAATTTTTTGTTGAAATTAGAAAAAAATAATAAAAAAACAAACCTATTTTCTTTCAGGTTGTTTTTGGGTTACAAAA
                                              1450                              1500
CAAAAGAAAACATGTTTCAAGGTACAATAATTATGGTTCTTTAGATTTCTGTAAAACAGCCTTATAGTTGGATTTATGACTATTAAAGTTAGTATACA
                                              1550                              1600
GCATACACAATCTATTCAAGGATATTTATAATGCAATTCCCTAAAAATAGTTTTGTATAACCAGTTCTTTTATCCGAACTGATACACGTATTTTAGCATA
                                              1650                              1700
ATTTTTAATGTATCTTCAAAAAACAGCTTCTGTGTCCTTTTCTATTAAACATATAAATTCTTTTTTATGTTATATATTTATAAAAGTTCTGTTTAAAAAGC
           .  <---------------   -35-----1750 ----------- -10---------->.         rbs  1800
CAAAAATAAATAATTATCTCTTTTTATTTATATTATATTGAAACTAAAGTTTATTAATTTCAATATAATAGAAATTTAATTTTATACAAAAAGGAGAACG
      -----> SIG                             1850                       -----> MAT
TATATGAAAAAACGAAAAGTGTTAATACCATTAATGGTATTGTCTACGATATTAGTTTCAAGCACAGGTAATTAGAGGTGATTCAGGCAGAAGTTAAAC
    MetLysLysArgLysValLeuIleProLeuMetAlaLeuSerThrIleLeuValSerSerThrGlyAsnLeuGluValIleGlnAlaGluValLysGln
                                              1950                              2000
AGGAGAACCGGTTATTAAATCAATCAGAATCAAGTCCCAGGGGTTACTAGGATACTATTTTAGTGATTGAATTTTCAAGCACCCATGGTGGTTACCTC
GluAsnArgLeuLeuAsnGluSerGluSerSerSerGlnGlyLeuLeuGlyTyrTyrPheSerAspLeuAsnPheGlnAlaProMetValValThrSer
                                              2050                              2100
TTCTACTACACGGGCATTTATCTATTCCTAGTTCTGAGTTACAAAATATTCCATCGGAAACCAATATTTTCAARCTGCTATTTGGTCAGGATTTATCAAA
SerThrThrGlyAspLeuSerIleProSerSerGluLeuGluAsnIleProSerGluAsnGlnTyrPheGlnSerAlaIleTrpSerGlyPheIleLys
```

```
                    2150              .              .              .        HindIII  2200
GTTAAGAAGAGTGCATGAATATACATTTGCTACTTCCGCTGATAATCATGTAACAATGTGGGTAGATGACCAAGAAGTCGATTAATAAAGCTTCTAATTCTA
ValLysLysSerAspGluTyrThrPheAlaThrSerAlaAspAsnHisValThrMetTrpValAspAspGlnGluValIleAsnLysAlaSerAsnSerAsn
                    2250              .              .              .              .        2300
ACAAAATCAGATTAGAAAAAGGAAGATTATATCAAATAAAAATTCAATATCAACGAGAAAATCCTACTCGAAAAAGGATTGGGATTTCAAGTTGTACTGGAC
LysIleArgLeuGluLysGlyArgLeuTyrGlnIleLysIleGlnTyrGlnArgGluAsnProThrGluLysGlyLeuAspPheLysLeuTyrTrpThr
                    2350              .              .              .              .        2400
CGATTCTCAAAATAAAAAAGAAGTGATTTCTAGTGATAACTTGCAATTGCCAGAATTAAAACAAAAATCTTCGAACTCAAGAAAAAAGCGAAGTACAAGT
AspSerGlnAsnLysLysGluValIleSerSerAspAsnLeuGlnLeuProGluLeuLysGlnLysSerSerAsnSerArgLysLysArgSerThrSer
                    2450              .              .              .              .        2500
GCTGGACCTACGGTTCCAGACCGTGACAATGATGCAATCCCTGATTCATTAGAGGTAGAAGGCATATACCGTTGATGTCAAAATAAAAGAACTTTTTCTTT
AlaGlyProThrValProAspArgAspAsnAspGlyIleProAspSerLeuGluValGluGlyTyrThrValAspValLysAsnLysArgThrPheLeuSer
                    2550              .              .              .              .        2600
CACCATGGATTTCTAATATTCATGAAAAGAAAGGCATTAACCAAATATAAATCATCTCCTGAAAAATGGAGTACGGCTTCTGATCCGTACAGTGATTCGA
ProTrpIleSerAsnIleHisGluLysLysGlyLeuThrLysTyrLysSerSerProGluLysTrpSerThrAlaSerAspProTyrSerAspPheGlu
                    2650              .              .              .              .        2700
AAAGGTTACAGGACCGGATTGATAAGAATGTATCACCAGAGGCAGAGCACCCCCTTGTGGCAGCTTATCCGATTGTACATGTAGATATGGAGAATATTATT
LysValThrGlyArgIleAspLysAsnValSerProGluAlaArgHisProLeuValAlaAlaTyrProIleValHisValAspMetGluAsnIleIle
                    2750              .              .              .              .        2800
CTCTCAAAAAATGAGGATCAATCCACACAGAATACTGATAGTGAACGAGAACAATAGTAAAAATACTTCTACAAGTAGGACACATACTAGTGAAGTAC
LeuSerLysAsnGluAspGlnSerThrGlnAsnThrAspSerGluThrArgThrIleSerLysAsnThrSerThrSerArgThrHisThrSerGluValHis
                    2850              .              .        .  EcoRI    .              2900
ATGGAAATGCAGAAGTGCATCGCGTCGTTCTTTGATATTGGTGGGAGTGTATCTGCAGGATTAGTAATTCGAATTCAAGTACGGTCGCAATTGATCATTC
GlyAsnAlaGluValHisAlaSerPhePheAspIleGlyGlySerValSerAlaGlyPheSerAsnSerAsnSerSerThrValAlaIleAspHisSer
                    2950              .              .              .              .        3000
ACTATCTCTAGCAGGGGAAAGAACTTGGGCTGAAACAATGGGTTTAAATACCGCTGATACAGCAAGATTAAAGCCAATATTAGATATGAAATACTGGG
LeuSerLeuAlaGlyGluArgThrTrpAlaGluThrMetGlyLeuAsnThrAlaAspThrAlaArgLeuAsnAlaAsnIleArgTyrValAsnThrGly
                    3050              .              .              .              .        3100
ACGGCTCCAATCTACAACGTGTTACCAACGACTTCGTTAGTGTTAGGAAAAAATCAAACACTCGCGACAATTAAAGCTAAGGAAAACCAATTAAGTCAAA
ThrAlaProIleTyrAsnValLeuProThrThrSerLeuValLeuGlyLysAsnGlnThrLeuAlaThrIleLysAlaLysGluAsnGlnLeuSerGlnIle
                    3150              .              .              .              .        3200
TACTTGCACCTAATAATTATTATCCTTCTAAAAACTTGGCGCCAATCGCATTAAATGCACAAGACGATTCAGTTCTACTCCAATTACAATGAATTACAA
LeuAlaProAsnAsnTyrTyrProSerLysAsnLeuAlaProIleAlaLeuAsnAlaGlnAspAspPheSerSerThrProIleThrMetAsnTyrAsn
                    3250              .              .              .              .        3300
TCAATTTCTTGAGTTAGAAAAAACGAAACAATTAAGATTAGATACGGATCAAGTATATGGGAATATAGCAACATACAATTTTGAAAATGGAAGAGTGAGG
GlnPheLeuGluLeuGluLysThrLysGlnLeuArgLeuAspThrAspGlnValTyrGlyAsnIleAlaThrTyrAsnPheGluAsnGlyArgValArg
                    3350              .              .              .              .        3400
GTGGATACAGGCTCGAACTGGAGTGAAGTGTTACCGCAAATTCAAGAACAACTGCACGTATCATTTTTAATGGAAAAGATTAAATCTGGTAGAAAGGC
ValAspThrGlySerAsnTrpSerGluValLeuProGlnIleGlnGluThrThrAlaArgIleIlePheAsnGlyLysAspLeuAsnLeuValGluArgArg
                    3450              .              .              .              .        3500
GGATAGCGGCGGTTAATCCTAGTGATCCATTAGAAACGACTAAACCGGATATGACATTAAAAGAAGCCCTTAAAATAGCATTTGGATTTAACGAACCGAA
IleAlaAlaValAsnProSerAspProLeuGluThrThrLysProAspMetThrLeuLysGluAlaLeuLysIleAlaPheGlyPheAsnGluProAsn
                    3550              .              .              .              .        3600
TGGAAACTTACAATATCAAGGGAAAGACATAACCGAATTTGATTTTAATTTCGATCAACAAACATCTCAAAATATCAAGAATCAGTTAGCGGAATTAAAC
GlyAsnLeuGlnTyrGlnGlyLysAspIleThrGluPheAspPheAsnPheAspGlnGlnThrSerGlnAsnIleLysAsnGlnLeuAlaGluLeuAsn
                    3650              .              .              .              .        3700
GCAACTAACATATATACTGTATTAGATAAAATCAAATTAAATGCAAAATGCAATATTTAATAGCGATAAACGTTTTCATTATGATAGAAATAACATAG
AlaThrAsnIleTyrThrValLeuAspLysIleLysLeuAsnAlaLysMetAsnIleLeuIleArgAspLysArgPheHisTyrAspArgAsnAsnIleAla
                    3750              .              .              .              .        3800
CAGTTGGGGCGGATGAGTCAGTAGTTAAGGAGGCTCATAGACAAGTAATTAATTCGTCAACAGAGGGATTATTGTTAAATATTGATAAGGATATAAGAAA
ValGlyAlaAspGluSerValValLysGluAlaHisArgGluValIleAsnSerSerThrGluGlyLeuLeuLeuAsnIleAspLysAspIleArgLys
```

```
                                          3850          .        .        .        .        3900
AATATTATCAGGTTATATTCTAGAAATTGAAGATACTGAAGGGCTTAAAGAAGTTATAAATGACAGATATGATATGTTGAATATTTCTAGTTTACGGCAA
IleLeuSerGlyTyrIleValGluIleGluAspThrGluGlyLeuLysGluValIleAsnAspArgTyrAspMetLeuAsnIleSerSerLeuArgGln
                                          3950          .        .        .        .        4000
GATGGAAAAACATTTATAGATTTTAAAAAATATAATGATAAATTACCGTTATATATAAGTAATCCTAATTATAAGGTAAATGTATATGCTGTTACTAAAG
AspGlyLysThrPheIleAspPheLysLysTyrAsnAspLysLeuProLeuTyrIleSerAsnProAsnTyrLysValAsnValTyrAlaValThrLysGlu
                                          4050          .        .        .        .        4100
AAAACACTATTATTAATCCTAGTGAGAATGGGGATACTAGTACCAACGGGATCAAGAAAATTTTAATCTTTTCTAAAAAAGGCTATGAGATAGGATAAGG
AsnThrIleIleAsnProSerGluAsnGlyAspThrSerThrAsnGlyIleLysLysIleLeuIlePheSerLysLysGlyTyrGluIleGly
                                      .   <-----4150------------    -------------------->  .
                                          4150          .        .        .        .        4200
TAATTCTAGGTGATTTTTAAATTATCTAAAAAACAGTAAAATTAAAACATACTCTTTTGTAGAAATCAAGGAGAGTATGTTTTAAACAGTAATCTAA
              BamHI                  .    4250        .        .        .        .        4300
ATCATCATAATCCTTTGAGATTGTTTGTAGGATCC
```

(A)

(B)

Figure. (A) and (B) show RNA/DNA secondary structure (stem-loop) diagrams.

(A)
```
            G     A
                A A
              A=T
              T=A
   4160 · G=C ·· 4170
              T=A
              T=A
              T-G
              T-G
              T=A
              C=G
              T=A
              C=G
              A=T
   4150 · T=A · 4180
              A=T
              C=G
              A=T
              A=T
              A=T
   4140  A=T 4190    4200
   4130  A=T T=A
AAACAGTAAAAT=AACAGTAATCTAA
```

(B)
```
            A   A
            T     T
              T=A
              A=T · 900
              T=A
              A=T
        890 · A=T
              A=T
              C=G
              A=T
              T=A
              C=G
              A=T
              A=T · 910
              A=T
              A=T
        880 · A=T
              A=T
              T=A
              C=G
              A=T
              A=T
              A=T
              T=A · 920
              A=T
              A=T
                   A
850   860   870 · A=T  930   940
TTTATCGGGTAAAATAAAA=TACCTCCTATTTTAT ·
              G   A
```

(D)

```
                              T   T
                   1750 . T
                              G   A
                              A=T
                              A=T
                          A   T=A
                          A
                          C
                              A=T
                              A=T
                              A=T · 1760
                   1740 . G=C
                              T=A
                              T=A
                              A=T
                              T=A
                              A=T
                              T=A
                              T=A
                              A=T
                              T=A · 1770
                   1730 . A=T
                              T=A
                              T=A
                              T=A
                              A=T
                          T   T
                          T   T
                   1720 T=A . 1780
         ÁTAATTATCTCTT=ÁTTTATÁCAAÁ
         1710                    1790
```

(C)

```
                  A   T
              C           A
                  G   G
      1300 . A=T
              A=T
              T=A
              T=A · 1310
              C=G
              T=A
              C=G
              C=G
              C=G
              C=G
          A
      1290 . A=T   T   A
              A         G
              A         G · 1320
              A=T
              A=T
              T           T
              T   G-T
              T=A
              T=A
          A       A
      1280 . A=T
              T
              T   T · 1330
              A=T
              A=T
              A=T
              C=G
              A=T
              A=T
              T-G
      1270 . T=A
          G   A
              T=A · 1340
              A=T
                  T
              T=A
              T-G
              T=A   1350
              A=T   1360
         TATATTAAAÁTGT=AAAÁTAATAAAAÁ
         1250  1260
```

Kyte-Doolittle Hydropathic Analysis of 764 Residue Polypeptide from PA

Polypeptide Molecular Weight = 83786 daltons
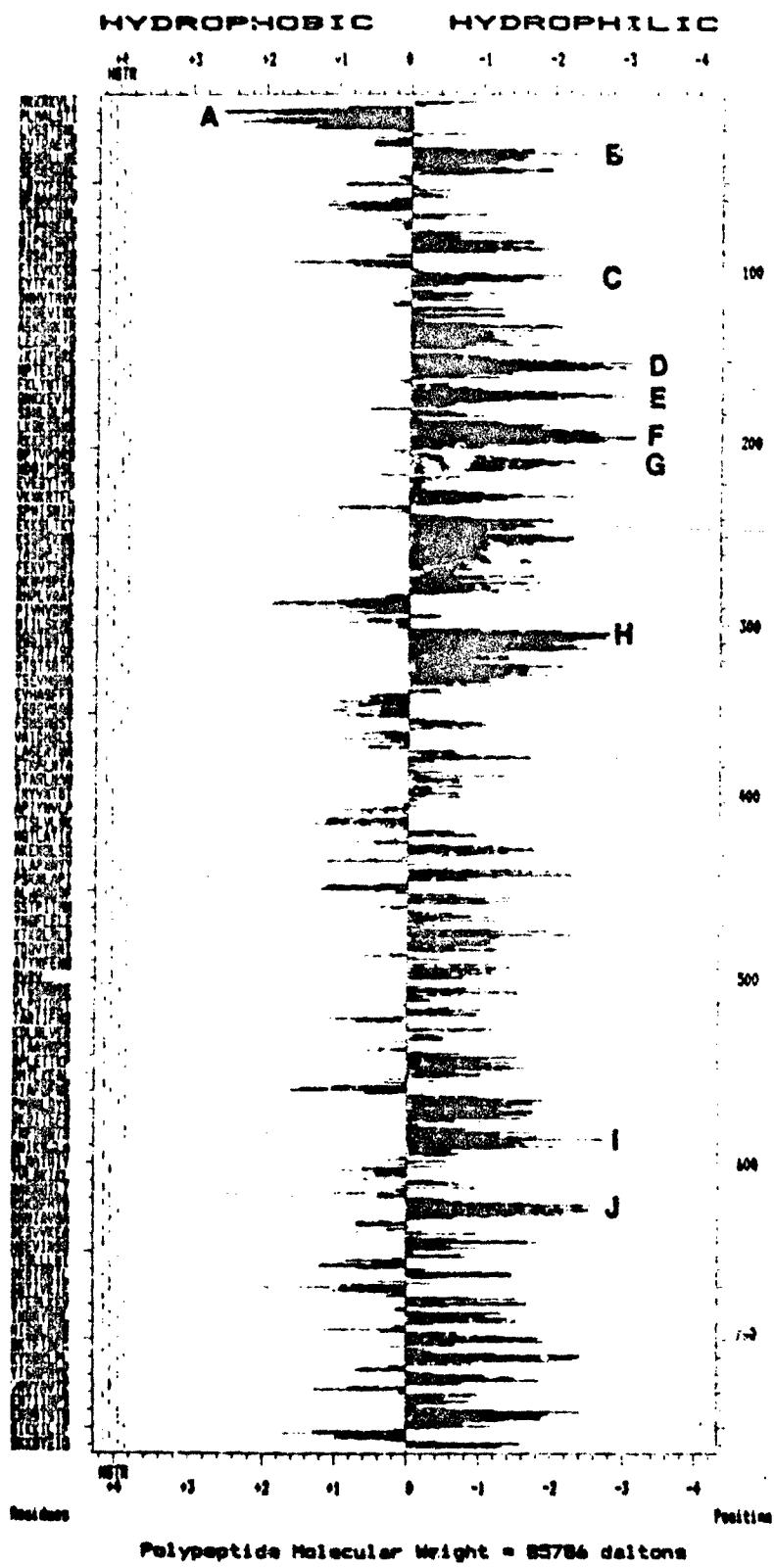
Kyte-Doolittle Hydropathic Analysis of 192 Residue Polypeptide from PA26ORF1

Polypeptide Molecular Weight = 21609 daltons
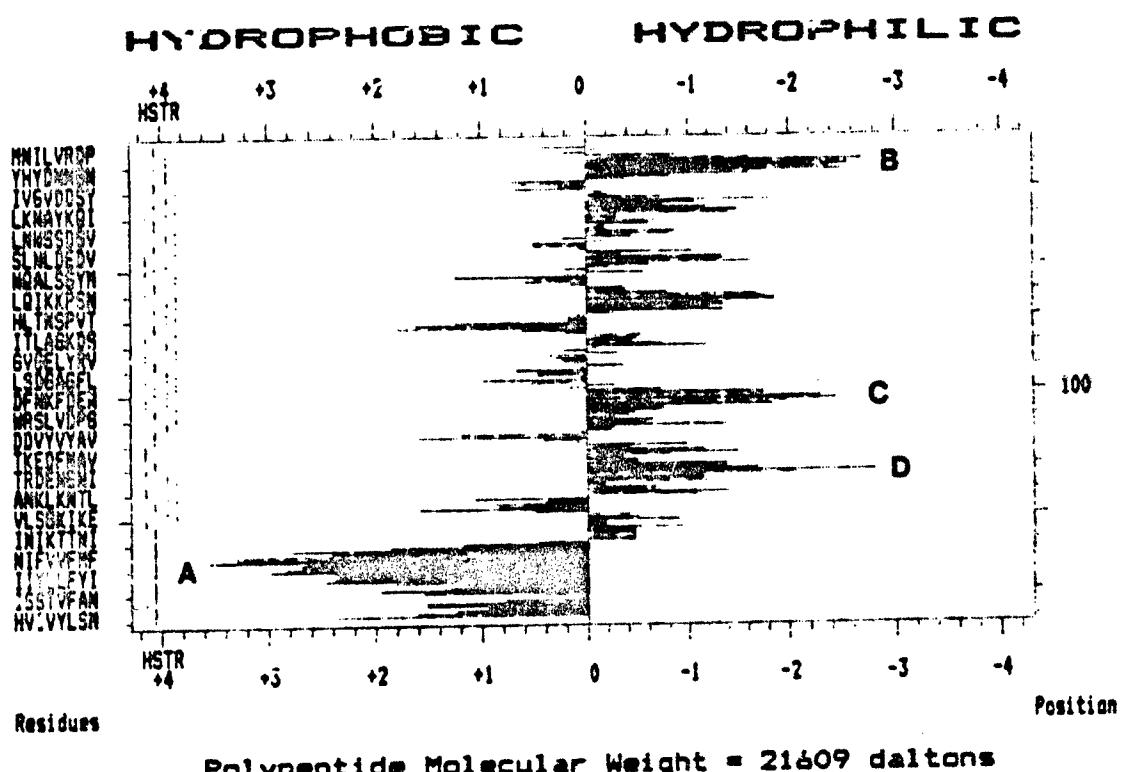
Table 1. Amino Acid Composition of PA and Codon Usage Comparisons[a]

	Codon	Species and Gene[b]								
		B.a. PA	B.a. ORF1	B.t. Cry Pro	C.t. TetTox	C.d. DTox	S.a. EntB	B.p. PTcxS3	E.c. ToxA	V.c. CTxA
Ala	GCU	31.7	12.5	31.3	37.5	42.2	42.9	4.0	28.6	23.1
41[a]	GCC	4.9	0.0	12.5	12.5	11.1	0.0	56.0	0.0	7.7
	GCA	46.3	75.0	32.8	41.7	20.0	57.1	8.0	50.0	53.8
	CCG	17.1	12.5	23.4	8.3	26.7	0.0	32.0	21.4	15.4
Arg	CGU	10.3	1.0	21.3	5.3	35.3	16.7	0.0	18.2	0.0
29	CGC	0.0	0.0	5.3	0.0	5.9	0.0	64.3	0.0	0.0
	CGA	10.3	50.0	13.3	5.3	11.8	33.3	0.0	4.5	0.0
	CGG	13.8	0.0	2.7	0.0	5.9	16.7	7.1	9.1	0.0
	AGA	55.2	50.0	44.0	57.9	17.6	33.3	7.1	54.5	33.3
	AGG	10.3	0.0	13.3	31.6	23.5	0.0	21.4	13.6	66.7
Asn	AAU	76.8	71.4	74.4	81.7	80.0	71.4	0.0	87.5	85.7
69	AAC	23.2	28.6	25.6	18.3	20.0	28.6	100.0	12.5	14.3
Asp	GAU	87.2	93.8	79.4	92.3	71.4	70.8	12.5	62.5	100.0
47	GAC	12.8	6.2	20.6	7.7	28.6	29.2	87.5	37.5	0.0
Cys	UGU	-	-	64.7	75.0	50.0	100.0	0.0	100.0	100.0
0	UGC	-	-	35.3	25.0	50.0	0.0	100.0	0.0	0.0
Gln	CAA	83.9	100.0	81.6	84.6	87.5	100.0	30.0	45.5	100.0
31	CAG	16.1	0.0	18.4	15.4	12.5	0.0	70.0	54.5	0.0
Glu	GAA	74.5	100.0	70.7	82.1	59.5	66.7	71.4	53.8	85.7
51	GAG	25.5	0.0	29.3	17.9	40.5	33.3	28.6	46.2	14.3
Gly	GGU	11.1	41.7	25.0	35.7	37.0	44.4	9.5	35.0	50.0
36	GGC	5.6	8.3	12.5	10.7	13.0	0.0	61.9	15.0	0.0
	GGA	52.8	41.6	45.0	50.0	21.7	44.4	14.3	45.0	50.0
	GGG	30.6	8.3	17.5	3.6	28.3	11.1	14.3	5.0	0.0
His	CAU	90.0	66.7	90.9	85.7	58.8	66.7	25.0	50.0	75.0
10	CAC	10.0	33.3	9.1	14.3	41.2	33.3	75.0	50.0	25.0
Ile	AUU	50.9	50.0	56.3	36.8	36.1	57.1	18.8	38.9	66.7
57	AUC	17.5	0.0	23.9	1.8	25.0	0.0	56.3	0.0	8.3
	AUA	31.6	50.0	19.7	61.4	38.9	42.9	25.0	61.1	25.0
Leu	UUA	67.7	52.6	45.0	60.4	20.0	52.4	0.0	61.1	33.3
62	UUG	12.9	5.3	11.0	13.2	15.0	23.8	12.5	5.6	11.1
	CUU	9.7	10.5	22.0	11.3	25.0	4.8	8.3	16.7	11.1
	CUC	3.2	5.3	2.0	0.0	10.0	4.8	20.8	5.6	0.0
	CUA	4.8	21.0	15.0	11.3	20.0	9.5	0.0	0.0	33.3
	CUG	1.6	5.3	7.0	3.8	10.0	4.8	58.3	11.1	11.1
Lys	AAA	78.3	75.0	72.7	92.0	72.5	76.5	0.0	83.3	63.6
60	AAG	21.7	25.0	27.3	8.0	27.5	23.5	100.0	16.7	36.4
Met	AUG	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
10										

Phe									
Phe	UUU	79.2	77.8	75.9	95.7	78.9	85.7	0.0	66.7	100.0
24	UUC	20.8	22.2	24.1	4.3	21.1	14.3	100.0	33.3	0.0
Pro	CCU	31.0	50.0	33.9	41.2	39.1	42.9	0.0	7.7	66.7
29	CCC	10.3	0.0	5.4	5.9	13.0	14.3	18.2	23.1	0.0
	CCA	37.9	50.0	42.9	47.1	26.1	42.9	18.2	53.8	33.3
	CCG	20.7	0.0	17.9	5.9	21.7	0.0	63.6	15.4	0.0
Ser	UCU	30.5	21.4	19.8	45.7	27.8	43.8	0.0	26.3	22.2
72	UCC	4.2	0.0	15.1	2.2	7.4	0.0	30.0	15.8	0.0
	UCA	20.8	35.7	27.9	23.9	13.0	12.5	0.0	31.6	33.3
	UCG	9.7	7.1	7.0	2.2	18.5	18.8	20.0	5.3	11.1
	AGU	31.9	14.3	22.1	17.4	13.0	18.8	0.0	10.5	33.3
	AGC	2.8	21.4	8.1	8.7	20.4	6.3	50.0	10.5	0.0
Thr	ACU	32.8	33.3	28.4	33.3	\).0	64.3	5.9	41.7	30.0
58	ACC	10.3	11.1	16.2	16.7	20.0	0.0	47.1	16.7	10.0
	ACA	37.9	55.6	31.1	50.0	23.3	14.3	5.9	41.7	40.0
	ACG	.9.0	0.0	24.3	0.0	16.7	21.4	41.2	0.0	20.0
Trp	UGG	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
7										
Tyr	UAU	78.6	81.8	76.9	88.2	72.2	77.3	42.1	73.9	100.0
28	UAC	21.4	18.2	23.1	11.8	27.8	22.7	57.9	26.1	0.0
Val	GUU	27.9	52.6	25.9	42.9	36.2	42.1	9.1	63.6	33.3
43	GUC	4.7	0.0	14.8	0.0	12.8	5.3	54.5	9.1	16.7
	GUA	39.5	42.1	40.7	50.0	31.9	31.6	18.2	18.2	50.0
	GUG	27.9	5.3	18.5	7.1	19.1	21.1	18.2	9.1	0.0
Nr		85787	21610	133047	65900	60753	31433	24989	29862	13909

ᵇThe following genes from the species listed were examined.

B.a. PA	=	*Bacillus anthracis* protective antigen gene
B.a. ORF1	=	*Bacillus anthracis* hypothetical protein gene 1 on pXO1 plasmid
B.t. Cry Pro	=	*Bacillus thuringiensis* crystal protein gene (39)
C.t. TetTox	=	*Clostridium tetani* tetanus toxin gene (6)
C.d. DTox	=	*Corynebacterium diphtheriae* diphtheria toxin gene (11)
S.a. EntB	=	*Staphylococcus aureus* enterotoxin B gene (16)
B.p. PToxS3	=	*Bordetella pertussis* pertussis toxin S3 binding subunit gene (24)
E.c. ToxA	=	*Escherichia coli* heat-labile enterotoxin A gene (45)
V.c. CTxA	=	*Vibrio cholerae* cholera toxin alfa subunit gene (28)

ᶜtotal number of specific amino acid residues deduced from protective antigen gene.